Extracción de Características en Audio con Redes Neuronales Convolucionales

García Mario Alejandro ✉[1], Rosset Ana Lorena[2], Eduardo Destéfanis[1]

[1]Universidad Tecnológica Nacional Facultad Regional Córdoba (UTN FRC)
[2]Universidad Nacional de Córdoba (UNC)
mgarcia@frc.utn.edu.ar

RESUMEN

La valoración de la calidad vocal mediante el análisis audio-perceptual es parte de la rutina clínica de evaluación de pacientes con trastornos de la voz. La debilidad de este método reside en la subjetividad y en la necesidad de que sea realizada por oyentes experimentados. Este proyecto tiene como objetivo la realización de una clasificación automática de la calidad vocal, valuada en la escala GRBAS, a través de características extraídas del análisis acústico de la señal y técnicas de aprendizaje automático. Particularmente, en este trabajo se muestran los resultados del diseño de las capas de extracción de características de una red neuronal profunda orientada a la clasificación de la calidad vocal.

Palabras clave: ***aprendizaje automático, aprendizaje profundo , análisis acústico***

CONTEXTO

Este trabajo de investigación se desarrolla en el marco del proyecto “Análisis acústico de la voz con técnicas de aprendizaje automático” (UTN3947) de la Universidad Tecnológica Nacional, Facultad Regional Córdoba y cuenta con la colaboración del Departamento de Investigación Científica, Extensión y Capacitación "Raquel Maurette", Escuela de Fonoaudiología, Facultad de Ciencias Médicas, Universidad Nacional de Córdoba.

1. INTRODUCCIÓN

Se intenta reconocer, de forma automática, características del análisis acústico de la voz que permitan clasificar muestras de audio. El estudio se enfoca en la medición de la calidad vocal según la escala GRBAS. La clasificación se realiza aplicando principalmente modelos de aprendizaje profundo (*deep learning)*, un subgrupo de técnicas del campo de aprendizaje automático (*machine learning*). Las grabaciones de la voz, la clasificación de los ejemplos y la validación de los resultados se realizan por especialistas en análisis de la voz de la Escuela de Fonoaudiología de la Universidad Nacional de Córdoba. El análisis acústico se lleva a cabo en conjunto (especialistas vocales e integrantes de UTN) y el modelado y desarrollo de los clasificadores por los integrantes de UTN.

GRBAS: La escala GRBAS es un método de valoración perceptivo-auditivo de la voz. Surge de la necesidad de estandarizar la valoración subjetiva y de interrelacionar los aspectos auditivos y fisiológicos de la producción vocal. Está basada en estudios del año 1966 de la *Japan Society of Logopedics and Phoniatrics* [1] y posteriormente divulgada y descripta por Minoru Hirano en el

año 1981 [2]. Consiste en la valoración de la fuente glótica a través de 5 parámetros que forman el acrónimo GRBAS:

G: (*Grade*) Grado general de disfonía.

R: (*Roughness*) Rugosidad, irregularidad de la onda glótica.

B: (*Breathiness*) Soplosidad, sensación de escape de aire en la voz.

A: (*Asteny*) Astenia, pérdida de potencia.

S: (*Strain*) Tensión, sensación de hiperfunción vocal.

Puede valorarse de dos maneras: a través de 4 grados, desde el 0 al 3 o mediante un valor en un rango continuo de 0 a 100. En ambas el 0 es ausencia de disfonía y el 3 o 100 implican disfonía severa. La escala fue mundialmente adoptada y validada en numerosos países [3-6]. Actualmente se utiliza en la investigación y de manera rutinaria en los consultorios de los profesionales que hacen clínica vocal. Sirve como metodología simple y al alcance de la mano para valorar la evolución pre-post tratamiento. La debilidad de este método reside en la subjetividad de la valoración de la voz y en la necesidad de que sea realizada por oyentes experimentados en la escucha y la disociación de los parámetros [7,8].

Análisis acústico: Existen otras formas de analizar la voz de manera más objetiva a través del análisis acústico. Éste consiste en la digitalización de la señal vocal y su análisis mediante gráficos como el Espectrograma, el espectro FFT (*Fast Fourier Transform*) o LPC (*Linear Predictive Coding*) y medidas numéricas de perturbación de la señal, como *Jitter, Shimmer* y HNR (*Harmonics to Noise Ratio*).

Para lograr una integración de la valoración subjetiva (GRBAS u otras escalas) con el análisis acústico, se han realizado numerosos trabajos de correlación [9,10], algunos relacionados a la voz normal y otros a diferentes patologías. Por ejemplo, el trabajo de Nuñez Batalla, F. et al [11] es un referente y establece una relación entre el parámetro de Astenia del GRBAS y el Espectrograma de banda angosta.

Aprendizaje automático: El aprendizaje automático o *machine learning* es un campo de las ciencias de la computación que abarca el estudio y la construcción de algoritmos capaces de aprender y hacer predicciones. Estas predicciones se pueden tomar como una clasificación de los datos de entrada a partir del reconocimiento de patrones existentes en los mismos

Estado del arte: La aplicación de técnicas de aprendizaje profundo es el estado del arte en el análisis automático de audio, con la detección de los fonemas pronunciados y la identificación de la persona que habla como objetivos principales [12-18], pero también utilizadas en detección de emociones, edad, género, etc.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

La línea de investigación que se presenta en este trabajo, se enfoca en el desarrollo de capas de extracción de características para un modelo de redes neuronales profundas.

La extracción de características es una de las etapas más importantes del aprendizaje automático. Esta tiene como objetivo la obtención de atributos del objeto a clasificar, de forma tal que un método de clasificación pueda encontrar regiones de decisión para cada una de las clases en el espacio formado por dichos atributos o características. En aprendizaje profundo, esta etapa se realiza en las primeras capas de la red neuronal.

Las medidas acústicas *shimmer*, *jitter* y HNR están relacionadas con la calidad vocal. A su vez, el cálculo de estas medidas depende del valor de la frecuencia fundamental de la voz F_0.

La estimación de F_0 es compleja y continúa siendo tema de investigación. Muchas de las técnicas de cálculo de F_0 parten de la información espectral de la señal. El espectro de frecuencias de una señal digital se obtiene con la Transformada Discreta de Fourier (DFT), por lo tanto, es deseable que la red neuronal sea capaz de calcular, en sus primeras capas, información equivalente a DFT del audio y la propague a las capas siguientes. Por ejemplo, la forma más frecuente de estimar F_0 implica el cálculo del cepstrum. Para calcular el cepstrum es necesario calcular dos veces sucesivas el *power spectrum* de la señal y esto a su vez implica dos DFT. Entonces, según el enfoque de nuestro trabajo, una red neuronal que calcule en sus primeras capas la DFT, recibirá en las capas posteriores la información necesaria para calcular F_0. A continuación se explica el cálculo de la DFT con una red neuronal.

Metodología:

Tanto la DFT como las redes neuronales, realizan transformaciones lineales. Por lo tanto, una red neuronal es capaz de calcular la DFT si sus pesos sinápticos son iguales a los coeficientes de la DFT [19].

Como para los objetivos del proyecto no es necesario preservar la información relativa a la fase en el espectro de frecuencias, se definen las capas necesarias para calcular la magnitud del espectro, es decir, el valor absoluto de la DFT. El valor absoluto de la DFT de una señal discreta $\boldsymbol{x}$ se define como:

$$|X_k| = \sqrt{\left(\sum_{n=0}^{N-1} x_n \phi(k,n)\right)^2 + \left(\sum_{n=0}^{N-1} x_n \psi(k,n)\right)^2}$$

donde N es la longitud de la secuencia $\boldsymbol{x}$, k es la frecuencia en el espectro,

$$\phi(k,n) = \cos(-2\pi kn/N),$$

$$\psi(k,n) = \sin(-2\pi kn/N).$$

La red neuronal que calcula la magnitud del espectro se muestra en la figura 1. Esta red devuelve un espectro de frecuencias para cada segmento de longitud N en el audio. En consecuencia a la salida de la red se obtiene un espectrograma.

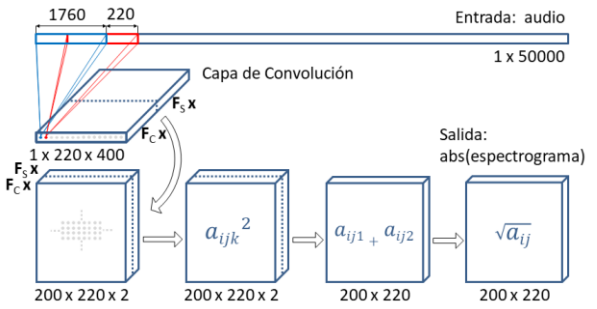

Figura 1. Modelo de neuronal de cálculo de la magnitud del espectrograma.

Los pesos de la capa de convolución son las matrices de tamaño $N \times N$, $\boldsymbol{F_C}$ y $\boldsymbol{F_S}$, que contienen los valores de $\phi(k,n)$ y $\psi(k,n)$ respectivamente. Más detalles se pueden ver en [19]. Con estos pesos, la red calcula el espectrograma de la misma forma que se hace con la DFT. La ventaja de hacerlo con una red neuronal, es que, si los pesos se pueden aprender, la red puede adaptarse a problemas particulares para los que exista un set de pesos mejores que los de $\phi(k,n)$ y $\psi(k,n)$. Para comprobar si la red tiene capacidad de adaptación, se realiza el entrenamiento partiendo de pesos aleatorios. Las entradas utilizadas son audios reales de vocales sostenidas de dos segundos de duración y las

salidas esperadas son espectrogramas calculados con la DFT.

3. RESULTADOS OBTENIDOS

A continuación se exponen los resultados de 30000 ciclos de entrenamiento del modelo propuesto. Los pesos se inicializaron con valores aleatorios entre -10^{-6} y 10^{-6} uniformemente distribuidos. Se utlizó el método de optimización Adam [20] con los parámetros provistos por los autores y las actualizaciones de los pesos se realizaron en batchs de tamaño 300 (la totalidad de datos de entrenamiento).

Los cálculos se realizaron sobre una GPU NVIDIA Titan Xp, donada a través del GPU Grant Program de NVIDIA.

El MSE de validación alcanzado fue 1.41×10^{-6} (9.79×10^{-6}% del valor medio de la salida esperada), mientras que para el mismo modelo con los pesos asignados de forma directa se logra un MSE < 10^{-9} [19].

Estos resultados indican que el modelo se puede entrenar y, por lo tanto, también puede adaptarse a problemas particulares.

4. FORMACIÓN DE RECURSOS HUMANOS

El equipo del proyecto está formado por un docente/investigador de la UTN FRC, dos docentes/investigadores de la UNC y cuatro alumnos de la carrera de grado de la UTN FRC.

Además de formación de los alumnos participantes, el conocimiento generado por el proyecto se incorporará a las cátedras de los docentes de la UTN y UNC.

5. REFERENCIAS

[1] Isshiki, N., Yanagihara, N., & Morimoto, M. (1966). *Approach to the objective diagnosis of hoarseness*. Folia Phoniatrica et Logopaedica, 18(6), 393-400.

[2] Hirano, M. (1981). *Clinical examination of voice* (Vol. 5). Springer.

[3] Yun, Y. S., Lee, E. K., Baek, C. H., & Son, Y. I. (2005). *The correlation of GRBAS scales and laryngeal stroboscopic findings for the assessment of voice therapy outcome in the patients with vocal nodules*. Korean Journal of Otolaryngology-Head and Neck Surgery, 48(12), 1501-1505.

[4] Hui, H., Weijia, K., & Shusheng, G. (2007). *The Validation of Acoustic Analysis and Subjective Judgment Scales of Several Voice Disorders* [J]. Journal of Audiology and Speech Pathology, 3, 010.

[5] Karnell, M. P., Melton, S. D., Childes, J. M., Coleman, T. C., Dailey, S. A., & Hoffman, H. T. (2007). *Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders*. Journal of Voice, 21(5), 576-590.

[6] Jesus, L. M., Barney, A., Couto, P. S., Vilarinho, H., & Correia, A. (2009, December). *Voice quality evaluation using cape-v and GRBAS in european Portuguese*. In MAVEBA (pp. 61-64).

[7] Kreiman, J., & Gerratt, B. R. (2010). *Perceptual assessment of voice quality: past, present, and future*. SIG 3 Perspectives on Voice and Voice Disorders, 20(2), 62-67.

[8] Núñez-Batalla et al (2012). El espectrograma de banda estrecha como ayuda para el aprendizaje del método GRABS de análisis perceptual de la disfonía. Acta

Otorrinolaringológica Española, 63(3), 173-179.

[9] Freitas, S. V., Pestana, P. M., Almeida, V., & Ferreira, A. (2015). *Integrating Voice Evaluation: Correlation Between Acoustic and Audio-Perceptual Measures*. Journal of Voice, 29(3), 390-e1.

[10] ELISEI, N. G. (2013). Percepción auditiva de voces patológicas. In XIV Reunión Nacional y III Encuentro Internacional de la Asociación Argentina de Ciencias del Comportamiento.

[11] Nuñez Batalla, F., Corte Santos, P., Señaris Gonzalez, B., Rodriguez Prado, N., Suárez Nieto, C. (2004) Evaluación espectral de la hipofunción vocal. Acta Otorrinolaringol. Esp. 55:327-333.

[12] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Kingsbury, B. (2012): Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, vol. 29.6, 82-97. IEEE.

[13] Mitra, V., Sivaraman, G., Nam, H., Espy-Wilson, C., Saltzman, E., Tiede, M. (2017) Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition. Speech Communication, vol. 89. pp 103-112.

[14] Collobert, R., Puhrsch, C., Synnaeve, G. (2016) Wav2letter: an end-to-end convnet-based speech recognition system.arXiv preprint arXiv:1609.03193.

[15] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Chen, J. (2016) Deep speech 2: End-to-end speech recognition in english and mandarin. International Conference on Machine Learning. pp. 173-182.

[16] Palaz, D., Collobert, R. (2015) Analysis of cnn-based speech recognition system using raw speech as input (No. EPFL-REPORT-210039). Idiap.

[17] Sainath, T. N., Kingsbury, B., Mohamed, A. R., Ramabhadran, B. (2013) Learning filter banks within a deep neural network framework. IEEE Workshop onASRU. pp 297-302. IEEE.

[18] Farrús, M. (2007) Jitter and shimmer measurements for speaker recognition. 8th Annual Conference of ISCA. pp. 778-781. (2007)

[19] García, M.A., Destéfanis, E.A.: Spectrogram Prediction with Neural Networks. XXIV Congreso Argentino de Ciencias de la Computación (2018).

[20] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)